DETECTING SEMANTIC DIFFERENCE: A NEW MODEL BASED ON KNOWLEDGE AND COLLOCATIONAL ASSOCIATION

Shiva Taslimipoor[1], Gloria Corpas[2], and Omid Rohanian[1]
[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
[2]Research Group in Computational Linguistics, University of Wolverhampton, UK and University of Malaga, Spain
{shiva.taslimi, omid.rohanian}@wlv.ac.uk
gcorpas@uma.es

Abstract

Semantic discrimination among concepts is a daily exercise for humans when using natural languages. For example, given the words, *airplane* and *car*, the word *flying* can easily be thought and used as an attribute to differentiate them. In this study, we propose a novel automatic approach to detect whether an attribute word represents the difference between two given words. We exploit a combination of knowledge-based and co-occurrence features (collocations) to capture the semantic difference between two words in relation to an attribute. The features are scores that are defined for each pair of words and an attribute, based on association measures, n-gram counts, word similarity, and Concept-Net relations. Based on these features we designed a system that run several experiments on a SemEval-2018 dataset. The experimental results indicate that the proposed model performs better, or at least comparable with, other systems evaluated on the same data for this task.

Keywords: semantic difference · collocation · association measures · n-gram counts · word2vec · Concept-Net relations · semantic modelling.

1. INTRODUCTION

Semantic modelling in natural language processing requires attending to both semantic similarity and difference. While similarity is well-researched in the community (Mihalcea & Hassan, 2017), the ability of systems in discriminating between words is an under-explored area (Krebs et al., 2018). A semantic model is more robust if it becomes sensitive to differences alongside similarities in meaning. For example, the concepts rain and snow are very similar, however the attribute liquid is challenging for a system whose aim is to understand natural languages. Such attributes (that might be very easy for humans to grasp) have been also researched as a kind of commonsense knowledge reasoning (Davis, 1990).

Distributional semantics, which is a common research area in semantic representation from the early times (cf. Firth, 1957 [1968]), is mainly exploited for finding similarities between words (Turney & Pantel, 2010). The main idea behind distributional semantics states that linguistic items with similar distributions have similar meanings (Blevins, 2016). The idea is further developed in state-of-the- art word representation models such as Mikolov et al. (2013). The effectiveness of a word representation model can be more rigorously evaluated by quantifying its strength in finding differences between words.

Santus et al. (2018) state that the task can also be a useful addition for the creation of ontologies and other types of lexical resources.

For this study, semantic difference is operationalised as follows. Given two semantically related words and a discriminative feature, the feature word should only characterise the first one. An example is the triplet *apple*, *banana*, *red*, in which *red* can only be an attribute for *apple* (i.e. the attribute *red* can discriminate *apple* from *banana*). In this sense, discriminative attributes are properties that people tend to find important for a given concept. The idea is that one can express semantic differences between concepts by referring to attributes of the concepts. This practice is defined by Krebs and Paperno (2016) as an evaluation set that captures differences between concepts. The so-called non-trivial semantic task was proposed by Krebs et al. (2018) as a competition in the Semantic Evaluation (SemEval 2018) conference. In such a competition, known as SemEval shared task, participants are provided with a shared annotated dataset and they are asked to design systems that automatically predict the annotation labels. After that, they are provided with a test dataset. All systems are then evaluated on the common dataset and compared to each other.

We propose two automatic approaches to capture discriminative attributes. One is a supervised support-vector machine (SVM) model (Cortes & Vapnik, 1995) and the other is a K-means clustering method (MacQueen, 1967). The features we design for both methods are scores computed for word pairs and triples with the aim of capturing different semantic relations. The first category of scores that we propose comes from co-occurrence statistics of the words. The motivation behind this is that the attribute discriminates a word, if its co-occurrence with the word is more salient compared to its co-occurrence with the second word. This lies at the heart of collocations (Cf. Smadja & McKeown, 1990; Hausmann, 2007). In this sense, an attribute is discriminative of a word with which it collocates. Two common features to extract this property are n-gram features and association measures which are further explained in Section 3.

Another related category of scores that we use comes from distributional similarity hypothesis. We expect that the attribute word should have a significantly higher similarity to the word that it discriminates compared to the other word. For this we use the recent word embedding methodology (Mikolov et al., 2013) which is widely adopted by state-of-the-art natural language processing systems.

The third category of scores is related to the hypothesis that discriminative attributes are common sense knowledge about a word. One promising resource to extract these knowledge-based features are semantic networks (Sowa, 1991) and we exploit ConceptNet (Speer & Havasi, 2013), in particular. In Section 3, we describe a formula that we propose to compute numerical features for each attribute word corresponding to an ordered pair of words. Our classification and clustering methodologies based on a knowledge-based ontology and co-occurrence counts are further evaluated and the results are reported and compared with other systems designed for the SemEval shared task. Our system ranked the fourth among the systems applied to the dataset of SemEval 2018 (Krebs et al., 2018). This study further elaborates on the advantages of the applied lexical features and discusses similarities and differences of the system with other systems that participated in the competition.

2. RELATED WORK

In the task of capturing discriminative attributes for words, different features have been used. These include collocational or co-occurrence-based features (Santus et al., 2018; Taslimipoor et al., 2018), word similarity features (Shiue et al., 2018), word embeddings (Santus et al., 2018), and finally the features extracted from taxonomy relations such as hypernymy (Is-A) or meronomy (Has-A) (Lai et al., 2018).

The term collocation was introduced by Firth (1957 [1968], 1968) to mean a mode of semantic analysis (meaning by collocation) and a stylistic means to characterise restricted languages. Later on collocation was equated with usual or habitual co-occurrence. Halliday’s redefinition of collocation in probabilistic terms marks the beginning of the distributional or statistical approach to collocation: "the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a*, *b*, *c* ..." (Halliday, 1966). The traditional lexico-semantic approach to collocation presupposes certain sense relations between the constituents of a collocation. Thus, collocations exhibit a bipartite structure, conventionally restricted, in which both collocates have a different semantic status: for example, in *commit suicide*, the base is the semantically autonomous word (*suicide*) and the verb *to commit* is the collocate, that is, the semantically dependent component (cf. Hausmann, 2007).

Word similarity and embeddings can all be grouped as distributional similarity features. The main idea of distributional similarity is that words that occur in the same contexts tend to have similar meanings. Distributional similarity can be approximated by different similarity measures between word vectors, including cosine, Jaccard coefficient, Euclidean distance, etc. (Lee, 1999). In this way, semantic difference can be modelled as the reverse order of similarity or can be judged based on the distributional similarity with a third word (Attia et al., 2018). However not all semantic differences can be adequately captured using this method. There are many cases where the difference between two words originates from the absence or the presence of a feature that cannot be directly mapped to the vector difference between two related words. One such example is *dolphin* and *narwhal* that only differ in having a *horn* (Krebs & Paperno, 2016). Such attribute is more visual and rarely occurs in text. Therefore, combining linguistic and conceptual information would potentially strengthen a semantic model in capturing the meaning of a word.

To tackle this issue, some studies rely on human annotated list of different attributes related to a concept which are called feature norms (McRae et al., 2005). Despite their strength in encoding semantic knowledge, feature norms have not been widely used in practice because they are usually small in size and require a lot of work to assemble (Fagarasan et al., 2015). Lazaridou et al. (2016) is an earlier attempt at identification of discriminative features which focuses on visual attributes.

The need for conceptual information also exists for systems that have to cope with commonsense reasoning such as question answering (Mcskimin, 1977) and word sense disambiguation (Sussna, 1993). This information can be obtained from manually or automatically created semantic networks such as BabelNet (Navigli & Ponzetto, 2012), ConceptNet (Speer & Havasi, 2013), etc. A semantic network is usually a directed or undirected graph structure consisting of nodes of concepts and edges which represent semantic relations between con cepts. ConceptNet is one such knowledge base including but not limited to relations such as *RelatedTo*, *IsA*, *HasA*, *PartOf*, *UsedFor* and *HasProperty*. Extracting any of these relations between a word and an attribute result in informative features to capture whether the attribute is discriminative of the word (Speer & Lowry-Duda, 2018). The representations learned on ConceptNet have also been proven successful in capturing discriminative attributes (Vinayan et al., 2019).

3. METHODOLOGY

Our goal is to define a simple interpretable metric that can be used to gauge semantic difference and to identify discriminative attributes. We hypothesise that for a triplet in this task, a stronger relation between the first word and the attribute (in comparison with the second word and the attribute)[1] is indicative of the attribute word being discriminative between the two words.

For each triple we define a discriminative score *Disc Score* (*w*1, *w*2, *attr*) as follows:

$$Disc\_Score\,(w1,\, w2,\, attr) = Score\,(w1,\, attr) - Score\,(w2,\, attr) \qquad (1)$$

[1] This stronger relation corresponds to more common semantic context and/or higher co-occurrence probability.

where *w*1, *w*2 and *attr* are the first, second, and third word respectively. *Score* is a variable function of relation between two words that can be any of the scores explained in Sections 3.1, 3.2, 3.3, and 3.4.

3.1 ASSOCIATION-BASED SCORE

Statistical association measures have a long history in language processing. With the availability of huge corpora, these measures can be even more effective than before in finding collocations and associations between words. Meaning by collocation is essentially a corpus-driven/corpus-based and distributional model of linguistic analysis which strives to statistically uncover significant word co-occurrences. This model also presupposes an underlying extraction method based on the analysis of discontinuous co-occurrences and word distance, where the units thus retrieved are also termed collocations or collocates (cf. Stubbs, 2002).

Collocational behaviour between two words is a strong signal that suggests that one of the words can identify the other. As an example, in the triplet (*hair, body, curly*), the association score in (*hair, curly*) is much higher than (*body, curly*), suggesting that *curly* is a discriminative attribute between the other two words. For each triplet in this task, collocational behaviour of the attribute word with the first two words is measured to see whether the first word can be a better collocate than the other. To this end, we use several different association measures[2] to compute the outputs of the *Score* function in Eq. 1.

We measure the association of two words based on their co-occurrence within a 5-word span. We use SketchEngine (Kilgarriff et al., 2004, 2014) to extract these statistics from the huge enTenTen corpus (Jakubíček, Kilgarriff, Kovář, Rychlý & Suchomel, 2013). Specifically, for each pair of words, we extract PMI (Church & Hanks, 1990) (known as MI in SketcEngine), MI3 (Oakes, 1998), log-likelihood (Dunning, 1993), T-score (Krenn & Evert, 2001), log-Dice (Dice, 1945), and Salience (Kilgarriff et al., 2014) all as defined in SketchEngine.

3.2 GOOGLE N-GRAMS

A second quantitative method to extract collocations is based on n-gram frequency analysis (continuous co-ocurrences) and it also requires very large data. In this case, the units of analysis are continuous sequences of two or more words which are retrieved from corpora according to a specified frequency threshold, regardless of their meanings (compositional or non-compositional) and their structural status (Stubbs, 2002). Unlike collocations, which can be discontinuous, n-grams are always a set of continuous co-occurring words. For instance, *excruciating pain* is an Adj. + N. collocation, but it does not constitute

---

[2] Statistical methods for the automatic extraction of collocations require large corpora and the use of an association measure or a combination of association measures: mutual information (MI), chi-square ($\chi 2$), phi-square ($\Phi 2$), log-likelihood (LR), etc. For a comprehensive list of association measures see Evert (2005).

one of the 3 bigrams that can be extracted from the sentence “The pain was excruciating” (1. *the pain*, 2. *pain was*, 3. *was excruciating*). This is a fundamental difference for their automatic extraction, as they require different techniques and procedures.

N-grams are frequently used in computational linguistics for a variety of purposes including language modelling and association measures based on lexical co-occurrence. A well-known collection of n-grams is Google Books Ngram Dataset.[3] This Dataset is a collection of phrases (between 1 and 5 words long) extracted from over 8 million books printed between 1500 and 2008.

We use PhraseFinder (Trenkmann, 2016), a free web API that makes it possible to look up words or phrases from this dataset using a wildcard-supporting query language. Using this resource, we derive two different features. In the first one, we only consider bigrams, and in the other, we consider up to 5-grams. In both cases, we count the number of times that words occur near one another within a given span, regardless of order. We follow the same formula as defined in Eq. 1. In order to eliminate the bias of high/low frequency words we divide *Disc_Score* by *Score* (*w*1, *attr*) + *Score* (*w*2, *attr*) that we compute from n-gram co-occurrence counts.

[3] https://books.google.com/ngrams.

3.3 WORD EMBEDDING-BASED SCORE

Word embedding is a type of word representation that allows words with similar meaning to be understood by machine learning algorithms: words are mapped into vectors of real numbers using a neural network.[4] The basic assumption is that this model can create vectors that categorise similar words together and place them far away from vector representation of different words. Thus, words that have the same meaning have a similar representation. For instance, word embedding will create the vector representation of drinks (*water*, *coffee*, *tea*, *juice*, *milk*, *wine*, etc.) as clearly separated from the vector of furniture (*table*, *cupboard*, *chair*, *bed*, *sofa*, *chest of drawers*, etc.).

In distributional semantics, word embeddings are used to induce meaning representations for words. These methods are inspired by neural network language modelling and have become a basic building block for most applications in computational linguistics. The most popular word embedding method is word2vec (with the skip-gram architecture) which learns dense vector representations for words using an unsupervised model. Word2vec's training objective is based on DH, defined so that the model can learn word vectors that are good at predicting nearby words (Mikolov et al., 2013). Another popular embedding technique is GloVe which, like word2vec, preserves semantic

---

[4] Word embedding is also termed distributed semantic model, distributed represented model or (semantic) vector space model.

analogies in the vector space. One major difference between the two models is that GloVe uses corpus statistics by training on global co-occurrence counts rather than local context windows (Pennington, Socher, & Manning, 2014).

In our system we use a concatenation of two sets of pre-trained embed- dings. The first is trained on English Wikipedia using a variation of word2vec (Bojanowski, Grave, Joulin, & Mikolov, 2016). The other called ConceptNet Numberbatch (Speer & Lowry-Duda, 2017), is an ensemble of pre-trained Glove and word2vec vectors whose values are readjusted using a technique called retrofitting (Faruqui et al., 2014). In retrofitting, the values of the embeddings are updated using a training function that considers relational knowledge.

Using each word embedding, we compute cosine similarity between each word in a triplet and the attribute word to account for the statistics *Score* (*w*1, *attr*) and *Score* (*w*2, *attr*) in Eq. 1.

3.4 CONCEPTNET SCORE

Co-occurrence based measures are not sufficient to account for all the various semantic relations that can exist between two words. Knowledge-based ontologies (e.g. ConceptNet, BabelNet etc) encode information about words and their relations in a structured way. This additional source of semantic information can be used to determine whether or not an attribute is discriminative. Because of its free web interface and ease of use, we use ConceptNet to empower our system with relational knowledge (Speer & Havasi, 2013). ConceptNet provides a large semantic graph to be used by computer applications. It represents general background knowledge and the way it is expressed in natural language (words and common phrases).

For any given ($w1$, $w2$, $attr$) triplet, using ConceptNet's REST API we query $w1$, limiting the number of search results to 1, 000. The output is a JSON file that contains all relations between the queried word and other concepts. We traverse all the relations and count the number of times $attr$ is linked to $w1$ to compute $score$ ($w1$, $attr$). We repeat the procedure for $w2$ and compute $score$ ($w2$, $attr$) and substitute them in Eq. 1.

Our goal is to define a simple interpretable metric so that we can gauge semantic difference and identify discriminative attributes. We hypothesise that for any triplet in this task, a stronger relation between the first word and the attribute (in comparison with the second word and the attribute) is indicative of the attribute word being discriminative between the two words.

4. EXPERIMENTS

4.1 DATA

The dataset provided by Krebs et al. (2018) as part of the shared task on "capturing discriminative attributes" (as explained in Section 1) contains manually verified triplets of the form <word1, word2, attribute>. The attribute characterises the first word only and hence based on this definition, semantic difference in this dataset is asymmetric (Krebs et al., 2018). The data includes both positive and negative examples. Positive examples are like <*tray*, *pan*, *rectangular*> and negative examples can range from the one that the attribute can refer to: none of the words (e.g. <*squirrel*, *leopard*, *fur*>) or both words (e.g. <*dresser*,*cupboard*,*large*>).

The triplets are divided into three sets: one set for training, a second set for validation and a third set for testing. The test set would be kept blind and the models are trained on training set and hyper-parameters are optimised on the validation set. In order to ensure that models do not rely on attribute memorisation, the division is done so that no attribute in the test set or the validation set is also present in the training set. The statistics about the dataset are represented in Table 1 from (Krebs et al., 2018).

	train	validation	test
positive	6591	1364	1047
negative	11191	1358	1293
total	17782	2722	2340

Table 1. The distribution of data into train, validation and test.

4.2 EXPERIMENTAL SETUP

The final feature set is the collection of Disc-Score measures based on the set of proposed scores. As a result we have six association-based scores, two google n-gram based scores, two embedding based scores, and one ConceptNet score. In total, we have eleven scores as our features.

In ConceptNet, reliability of each relation is given by a weight score. We decided to ignore this information and opted for raw counts because it did not help performance. Furthermore, binarising the scores based on raw counts (with 0 as a threshold) slightly improved the results. We use the features in both a supervised scenario (using SVM) and an unsupervised scenario (using KMeans). In both cases, all eleven features are exploited.

4.3 EVALUATION METRICS

The evaluation in this shared task is in terms of the average of positive and negative F1-scores which are standard in binary classification tasks. In this chapter, we report the precision, recall and F1-score for both positive and negative labels separately, along with the average F1-score.

The baseline system adopted by Krebs et al. (2018) is a simple unsupervised method that classifies a triplet as positive if the similarity of the attribute and the first word is greater than its similarity to the second word. The performance of the baseline is reported in Section 5. They also calculate the upper bound performance by human on the dataset which is F1-score of 0.9.

5. RESULTS AND DISCUSSION

Table 2 shows the results on both validation and test sets. The validation set is available to the system at the time of training and we perform hyperparameter optimisation on that. The test set however is blind to the system. We report the results of both our supervised (SVM) and unsupervised (KMeans) models and compare them with the baseline and the top system (Lai et al., 2018) applied to this dataset.

			Precision	Recall	F1-score	Average F1-score
Validation	SVM	pos	0.7679	0.5652	0.6512	0.6913
		neg	0.6548	0.8284	0.7315	
	KMeans	pos	0.7039	0.6833	0.6935	**0.6972**
		neg	0.6910	0.7113	0.7010	
TEST	baseline		-	-	-	0.607
	SVM	pos	0.7299	0.6065	0.6625	**0.7142**
		neg	0.7197	0.8183	0.7658	
	KMeans	pos	0.6464	0.7001	0.6722	0.6930
		neg	0.7396	0.6899	0.7139	
	Top System 1	-	-	-	-	0.75

Table 2. Results on Validation and TEST sets.

According to Table 2, our systems significantly outperform the baseline and underperform the Top System 1 by lower F1-score of less than 0.04. It is surprising that the unsupervised model (KMeans) can cluster the validation data as well as or even better than the supervised classification approach (SVM). Unsupervised models do not require training data. These models only use the validation data for hyperparameter optimisation.

This can be explained by the fact that the features we employ for this task are all computed using a formula that is specifically defined to represent semantic difference, and that finding whether a feature is discriminative between two words closely correlates with the semantic difference between them. Another reason could be that the training dataset is very noisy (cf. Krebs et al., 2018). The best performing system (Top System 1), in fact, got the best result by being trained directly on the validation data, otherwise by training on both train and validation data, their performance was reported to be 0.721 (Lai et al., 2018). This system is similar to our system in the sense that they are using SVM and word similarities as one of their feature types. One difference is that they rely on taxonomy relations from Probase, which can be considered a limitation when such taxonomies are not available.

We can see from the results that our features are well generalised as they lead to even better performance on the held-out test data. In order to see the effectiveness of the scores we obtained from ConceptNet, we retrained the model excluding the ConceptNet based measure and also the vectors derived from Numberbatch embedding. As a result, the validation performance dropped to 0.6857 and the test result decreased to 0.6969 in terms of average F1-score.

One advantage of our system is its ability to capture genealogical and kinship relations, as in (*grandson, brother, male*). Some train and test triplets require hierarchical reasoning, as in (*invertebrate, insect, shell*). Our model captures these kinds of relations very well, as it has access to information from a knowledge base. It is worth noting that a large part of the test triplets requires the knowledge to understand whether something is a constituent of another entity, as in (*beer, wine, foam*). It appears that these relations are well captured using co-occurrence-based metrics (collocations) alone since deleting knowledge-based features leaves the results for these triplets for the most part unchanged.

6. CONCLUSIONS

Semantic similarity is a well-represented research topic in Computational Linguistics. There are plenty of procedures and metrics to compute semantic similarity among words or even texts that use statistics from corpora. In this paper we have described an alternative procedure from the opposite perspective: computing semantic difference. Our model provides a simple metric in order to discriminate among words in relation to an attribute. The approach is based on a combination of knowledge-based and co-occurrence features (collocations, n-grams and word embeddings). Simple and robust, our method can be successfully used as an addition to semantic modelling, as it computes the difference among words optimally.

References

Attia, M., Samih, Y., Faruqui, M., & Maier, W. (2018, June). GHH at SemEval-2018 task 10: Discovering discriminative attributes in distributional semantics. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 947–952). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S18-1155 doi: https://doi.org/10.18653/v1/S18-1155

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford: Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16* (1), 22–29.

Cortes, C., & Vapnik, V. (1995, Sep 01). Support-vector networks. *Machine Learning*, *20* (3), 273–297. Retrieved from https://doi.org/10.1007/BF00994018 doi: https://doi.org/10.1007/BF00994018

Davis, E. (1990). *Representations of commonsense knowledge*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297–302.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19* (1), 61–74.

Evert, S. (2005). *The statistics of word co-occurrences : word pairs and collocations* (Doctoral dissertation, Institut fr maschinelle Sprachverarbeitung, University of Stuttgart, Holzgartenstr. 16, 70174 Stuttgart). Retrieved from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371.

Fagarasan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th international conference on computational semantics* (pp. 52–57).

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*

Firth, J. (1957 [1968]). A synopsis of linguistic theory, 1930-1955. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firth (1952-59)* London: Longmans. (pp. 168-205).

Firth, J. (1968). Linguistic analysis as a study of meaning. In F. R. Palmer (Ed.), *Selected Papers of J. R. Firt*h (1952-59) London: Longmans. (pp. 12–26).

Halliday, M. (1966). Lexis as a linguistic level. In Bazell, C.E., Catford, J.C., Halliday, M.A.K., Robins, R.H. (Eds.). *In memory of John Firth* London: Longman. (pp. 148–162).

Hausmann, F. (2007). Die Kollokationen im Rahmen der phraseologie systematische und historische Darstellung. *Zeitschrift Fr Anglistik und Amerikanistik*, *55*, 217–234.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., & Suchomel, V. (2013). The tenten corpus family. In *7th international corpus linguistics conference cl* (p. 125-127).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Suchomel, V. (2014, Jul 01). The Sketch engine: ten years on. *Lexicography*, *1* (1), 7–36. Retrieved from https://doi.org/10.1007/s40607-014-0009-9 doi: https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlỳ, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress,* edited by G. Williams, G and S. Vessier, 105-116. Lorient: Université de Bretagne-Sud.

Krebs, A., Lenci, A., & Paperno, D. (2018). SemEval-2018 task 10: Capturing discriminative attributes. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 732–740). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S18-1117 doi: https://doi.org/10.18653/v1/S18-1117

Krebs, A., & Paperno, D. (2016). Capturing discriminative attributes in a distributional space: Task proposal. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP* (pp. 51–54).

Krenn, B., & Evert, S. (2001). Can we do better than frequency? a case study on extracting pp-verb collocations. *Proceedings of the ACL Workshop on Collocations*, 39–46.

Lai, S., Leung, K. S., & Leung, Y. (2018, June). SUNNYNLP at SemEval-2018 task 10: A support-vector-machine-based method for detecting semantic difference using taxonomy and word embedding features. In *Proceedings of the 12th international workshop on semantic evaluation*, 741–746. New Orleans, Louisiana: Association for Computational Linguistics.

Lazaridou, A., Baroni, M., et al. (2016). The red one!: On learning to refer to things based on discriminative properties. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 213–218).

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (pp. 25–32). Stroudsburg, PA, USA: Association for

Computational Linguistics. Retrieved from https://doi.org/10.3115/1034678.1034693
doi: https://doi.org/10.3115/1034678.1034693

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 281– 297). Berkeley, Calif.: University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200512992

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37* (4), 547–559.

Mcskimin, J. R. (1977). The use of a semantic network in a deductive question-answering system. In *Proc. IJCAI 5* (pp. 50–58).

Mihalcea, R. & Hassan, S. (2017). Similarity. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217 - 250.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Santus, E., Biemann, C., & Chersoni, E. (2018, June). BomJi at SemEval-2018 task 10: Combining vector-, pattern- and graph-based information to identify discriminative attributes. In *Proceedings of the 12th international workshop* *on semantic evaluation* (pp. 990–994). New Orleans, Louisiana: Association for Computational Linguistics.

Shiue, Y.-T., Huang, H.-H., & Chen, H.-H. (2018, June). NTU NLP lab system at SemEval-2018 task 10: Verifying semantic differences by integrating distributional information and expert knowledge. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1027–1033). New Orleans, Louisiana: Association for Computational Linguistics.

Smadja, F. A., & McKeown, K. R. (1990, June). Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th annual meeting of the association for computational linguistics* (pp. 252–259). Pittsburgh, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P90-1032 doi: https://doi.org/10.3115/981823.981855

Sowa, J. F. (Ed.) (1991). *Principles of Semantic Networks. Explorations in the Representation of Knowledge*. San Mateo, California: Morgan Kaufmann.

Speer, R., & Havasi, C. (2013). Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web meets NLP* (pp. 161–176). Springer.

Speer, R., & Lowry-Duda, J. (2017). Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th international workshop on semantic evaluation (semeval- 2017)* (pp. 85–89).

Speer, R., & Lowry-Duda, J. (2018, June). Luminoso at SemEval-2018 task 10: Distinguishing attributes using text corpora and relational knowledge. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 985–989). New Orleans, Louisiana: Association for Computational Linguistics.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, *7* (2), 215–244.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the second international* *conference on information and knowledge management* (pp. 67–74). New York, NY, USA: ACM.

Taslimipoor, S., Rohanian, O., Ha, L. A., Corpas Pastor, G. & Mitkov, R. (2018, June). Wolves at SemEval-2018 task 10: Semantic discrimination based on knowledge and association. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 972–976). New Orleans, Louisiana: Association for Computational Linguistics.

Trenkmann, M. (2016). *PhraseFinder – Search millions of books for language use.* http://phrasefinder.io/. (Accessed: 2018-01-30)

Turney, P. D., & Pantel, P. (2010, January). From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, *37* (1), 141–188. Retrieved from http://dl.acm.org/citation.cfm?id=1861751.1861756

Vinayan, V., Anand Kumar, M., & Soman, K. P. (2019). Capturing discriminative attributes using convolution neural network over conceptnet number-batch embedding. In V. Sridhar, M. Padma, & K. R. Rao (Eds.), *Emerging research in electronics, computer science and technology* (pp. 793–802). Singapore: Springer Singapore.